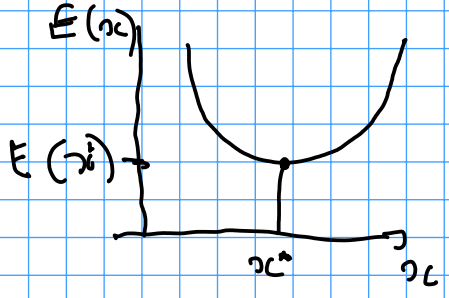
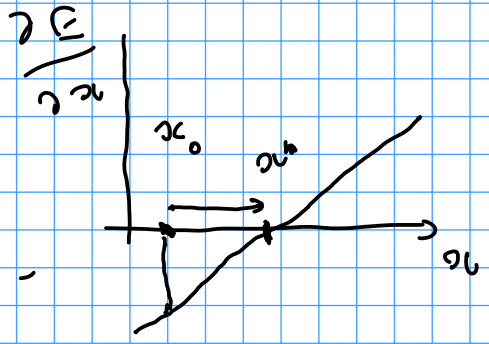


$$E(x) = \frac{1}{2} q (x - x^*)^2 + E(x^*)$$

$$\frac{\partial E(x)}{\partial x} = q (x - x^*)$$

$$\frac{\partial^2 E(x)}{\partial x^2} = q$$



$$\left(\frac{\partial E / \partial x (x^0)}{x^0 - x^*} \right) = q = \frac{\partial^2 E}{\partial x^2}(x_0)$$

$$x^* = x^0 - \left(\frac{\partial^2 E_{x_0}}{\partial x^2} \right)^{-1} \frac{\partial E}{\partial x}(x_0)$$

$$E(w) = \frac{1}{P} \sum_{p=1}^{P} \frac{1}{2} \left(y^p - w^T x^p \right)^2$$

$$\frac{\partial E}{\partial w} = -\frac{1}{P} \sum_{p=1}^{P} \left(y^p - w^T x^p \right) \cdot x^{pT}$$

$$H = \frac{\partial^2 E}{\partial w \, \partial w^T} = \frac{1}{P} \sum_{p=1}^{P} x^p \cdot x^{pT}$$

Covariance matrix of samples

Hessian matrix

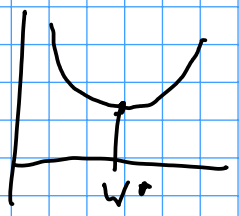
$$w_{t+1} = w_t - H^{-1}_{(w_t)} \frac{\partial E}{\partial w}(w_t)$$

Solve for w_{t+1}:

$$H \left(w_{t+1} - w_t \right) = \frac{\partial E}{\partial w}(w_t)$$

/

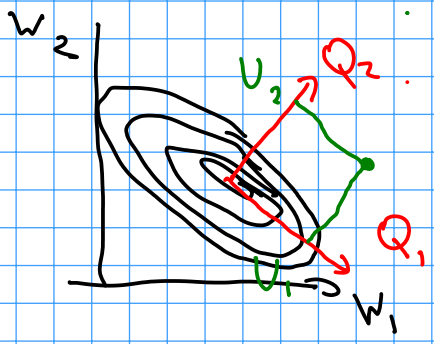
$$E(w) = E(w^*) + \frac{1}{2}(w-w^*)^T H (w-w^*)$$



$$\| \quad = \quad \| \quad + \frac{1}{2}(w-w^*)^T Q^T \Lambda Q (w-w^*)$$

$$v = Qw$$

$$E(v) = E(v^*) + \frac{1}{2}(v-v^*)^T \Lambda (v-v^*)$$



$$v_{k+1} = v_k - \underbrace{\Lambda^{-1}}_{H^{-1}} \underbrace{\left[\Lambda (v_t - v^*)\right]}_{\text{gradient}} \rightarrow v_{t+1} = v^*$$

$$v_{k+1}(i) = v_{k(i)} - \underbrace{\lambda_i^{-1}}\left[\frac{\partial E}{\partial v}(i)\right]$$

$$w_i \leftarrow w_i - \eta_i \left(\frac{\partial E}{\partial w}\right)_i$$

$$\eta_i = \frac{1}{\left[\frac{\partial^2 E}{(\partial w_i)^2}\right]^+ + \gamma}$$

diagonal Newton

Levenberg-Marquardt

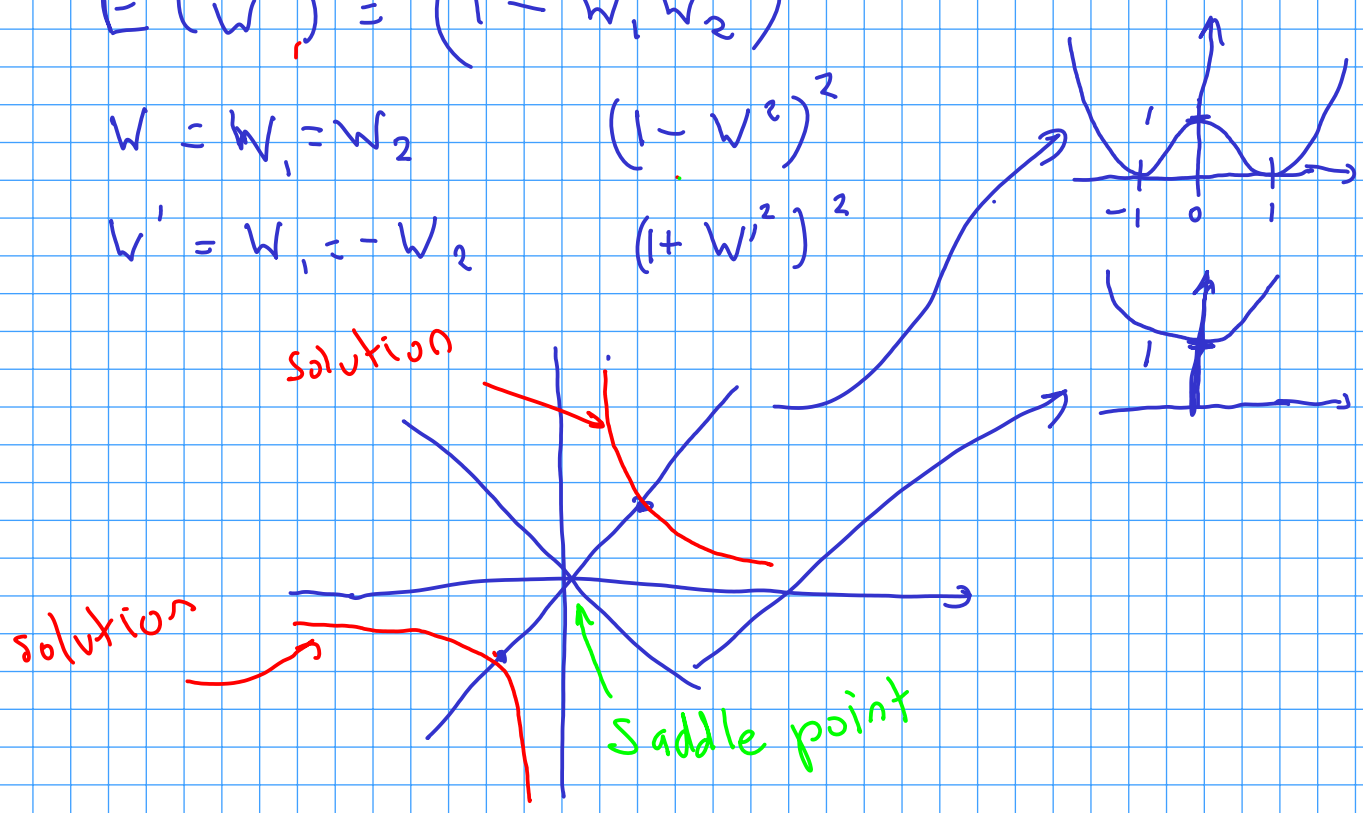
$$(H + \gamma I)^{-1} \frac{\partial E}{\partial w}$$

positive semi-definite estimate of Hess

$$E(w) = (1 - w_1 w_2)^2$$

$w = w_1 = w_2$ $\quad (1 - w^2)^2$

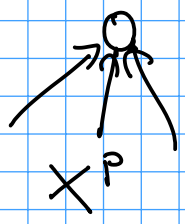
$w' = w_1 = -w_2$ $\quad (1 + w'^2)^2$



$$M_{t+1} = M_t + \alpha \frac{\partial \epsilon}{\partial W}$$

Nesterov momentum

$$W_{t+1} = W_t - \eta M_t$$



$$H = \frac{1}{P} \sum_{p=1}^{P} x^p x^{pT}$$

speed $H = Q^T \Lambda Q$

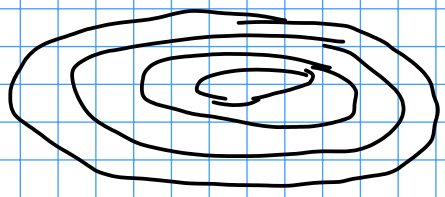
$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_N \end{bmatrix}$$

eigenvalues of H

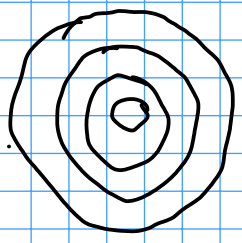
eigenvectors of H

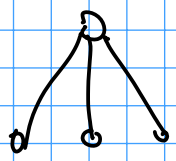
When H is diagonal

principal axes are aligned with frame of reference



$$\begin{bmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{bmatrix}$$





$$H = \frac{1}{P} \sum_{p=1}^{P} x^p x^{p^T} = \lambda I \qquad H_{ij} = \sum_{p=1}^{P} x_i^p x_j^p$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$$

x_i : - zero mean
- variance 1
- decorrelated

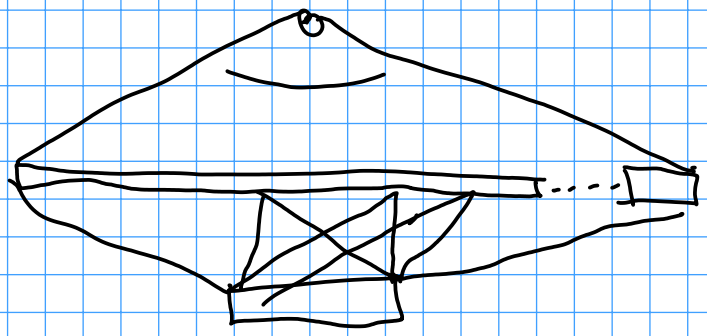
$X^p \longrightarrow Z^p$ 0 mean, variance 1, uncorrelated

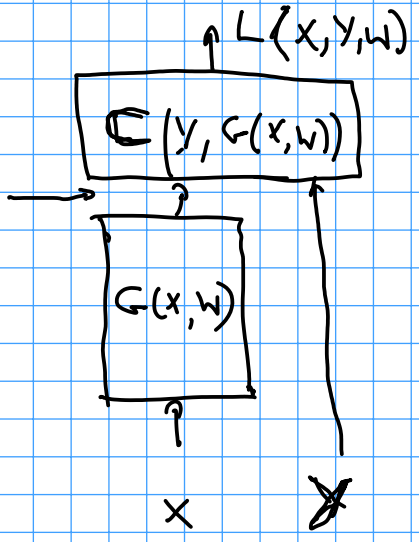
↑ whitening

Karhunen-Loève transform

X $\quad H = Q^T \Lambda Q \quad Z = \Lambda^{-1} Q(X - \bar{X})$

$H = \sum_p (x^p - \bar{x}^p)(x^p - \bar{x}^p)^T$





Supervised learning

C differentiable
G differentiable
y observed
X observed
L(x,y,w) computed

↕

SGD Backprop

C non diff
G diff / G non diff
y obs
X obs
L computed

↓

black-box optim
REINFORCE

BL : Y is not observed
L is observed
C unknown